Measuring the effective complexity of cosmological models

Martin Kunz,[1, *] Roberto Trotta,[2, †] and David R. Parkinson[3, ‡]

[1]*Département de Physique Théorique, Université de Genève,*
24 quai Ernest Ansermet, CH-1211 Geneva 4, Switzerland
[2]*Oxford University, Astrophysics, Denys Wilkinson Building, Keble Road, OX1 3RH, United Kingdom*
[3]*Astronomy Centre, University of Sussex, Brighton, BN1 9QH, United Kingdom*
(Dated: February 3, 2008)

We introduce a statistical measure of the effective model complexity, called the *Bayesian complexity*. We demonstrate that the Bayesian complexity can be used to assess how many effective parameters a set of data can support and that it is a useful complement to the model likelihood (the evidence) in model selection questions. We apply this approach to recent measurements of cosmic microwave background anisotropies combined with the Hubble Space Telescope measurement of the Hubble parameter. Using mildly non–informative priors, we show how the 3-year WMAP data improves on the first-year data by being able to measure both the spectral index and the reionization epoch at the same time. We also find that a non-zero curvature is strongly disfavored. We conclude that although current data could constrain at least seven effective parameters, only six of them are required in a scheme based on the ΛCDM concordance cosmology.

PACS numbers: 98.80.Es, 02.50.-r, 98.70.Vc

I. INTRODUCTION

The quest for a cosmological standard model is being driven by an increasing amount of high quality observations. An important and natural question concerns the number of fundamental parameters of the underlying physical model. How many numbers are necessary to characterize the Universe? Or in other words, how complex is the Universe? Generally the term "complexity" is employed in a rather loose fashion: a more complex model is one with a larger number of parameters that can be adjusted over a large range to fit the model to the observations. In this paper we will try to measure the effective number of model parameters that a given set of data can support. Because of the connection to the data, we will call this the *effective complexity* or *Bayesian complexity*. Our main purpose is to present a statistically sound quantity that embodies in a quantitative way the above notion of complexity when a model is compared to the observations.

Bayesian model comparison makes use of an Occam's razor argument to rank models in term of their quality of fit *and* economy in the number of free parameters. A model with more free parameters will naturally fit the observations better, but it will also be penalized for the wasted parameter space that the larger number of parameters implies. Several studies have made use of Bayesian model comparison to assess the viability of different models in the cosmological context [1, 2, 3]. In this work we show that Bayesian complexity is an ideal complement to model selection in that it allows to identify the number of effective parameters supported by the data.

We start by introducing our notation and the fundamentals of Bayesian statistics and model comparison. We then present the Bayesian complexity and illustrate its use in the context of a toy model in section III. In section IV we apply it to observations of cosmic microwave background anisotropies and we conclude in section V.

II. BAYESIAN MODEL SELECTION AND COMPLEXITY

A. Model comparison

We first briefly review the basic ingredients of Bayesian statistics and some relevant aspects of information theory. This serves both to introduce our notation and to remind the reader of the main points. We will use a fairly compact notation where possible and refer the reader to *e.g.* [4] for the exact mathematical definitions. Specifically, for an outcome x of a random variable X we will write $p(x)$ for the probability distribution function (pdf), ie the probability that X takes a certain value x. In the case of a multi-dimensional parameter space we will write $p(x)$ as a short form of the joint probability over all components of x, $p(x_1, x_2, \ldots, x_n)$. The conditional probability of x given y is written $p(x|y)$.

The starting point of our analysis is *Bayes theorem*:

$$p(x|y, I) = \frac{p(y|x, I)p(x|I)}{p(y|I)}. \quad (1)$$

Here, the quantity I represents a collection of all external hypotheses and our model assumptions.

Given the data d and a model $\mathcal{M}$ with n free parameters θ, statistical inference deals with the task of determining a posterior pdf for the free parameters of the

*Electronic address: Martin.Kunz@physics.unige.ch
†Electronic address: rxt@astro.ox.ac.uk
‡Electronic address: D.R.Parkinson@sussex.ac.uk

model, $p(\theta|d, \mathcal{M})$. The latter is computed via Bayes theorem as

$$p(\theta|d, \mathcal{M}) = \frac{p(d|\theta, \mathcal{M})p(\theta|\mathcal{M})}{p(d|\mathcal{M})}. \qquad (2)$$

On the right–hand side of this equation, $p(d|\theta, \mathcal{M})$ is the probability of obtaining the observed data given the parameter value θ. Since the observed data are a fixed quantity, we interpret $p(d|\theta, \mathcal{M})$ as a function of θ and we call it the *likelihood*, $\mathcal{L}(\theta) \equiv p(d|\theta, \mathcal{M})$. The *prior pdf* $p(\theta|\mathcal{M})$ embodies our state of knowledge about the values of the parameters of the model before we see the data. There are two conceptually different approaches to the definition of the prior: the first takes it to be the outcome of previous observations (*i.e.*, the posterior of a previous experiment becomes the prior for the next), and is useful when updating ones knowledge about the values of the parameters from subsequent observations. For the scope of this paper, it is more appropriate to interpret the prior as the available parameter space under the model, which then is collapsed to the posterior after arrival of the data. Thus the prior constitutes an integral part of our model specification. In order to avoid cluttering the notation, we will write $\pi(\theta) = p(\theta|\mathcal{M})$, with the model dependence understood whenever no confusion is likely to arise.

The expression in the denominator of (2) is a normalization constant and can be computed by integrating over the parameters,

$$p(d|\mathcal{M}) = \int d\theta \, \pi(\theta)\mathcal{L}(\theta). \qquad (3)$$

This corresponds to the average of the likelihood function under the prior and it is the fundamental quantity for model comparison. The quantity $p(d|\mathcal{M})$ is called *marginal likelihood* (because the model parameters have been marginalized), or in recent papers in the cosmological context, the *evidence.* In the following we shall refer to it as to the *likelihood for the model* [13]. The posterior probability of the model is then, using Bayes theorem again,

$$p(\mathcal{M}|d) = \frac{p(d|\mathcal{M})\pi(\mathcal{M})}{p(d)}, \qquad (4)$$

where $\pi(\mathcal{M})$ is the prior for the model. The quantity in the denominator on the right–hand side is just a normalization factor depending on the data alone, which we can ignore. When comparing two models, $\mathcal{M}_1$ versus $\mathcal{M}_2$, one introduces the *Bayes factor* B_{12}, defined as the ratio of the model likelihoods

$$\frac{p(\mathcal{M}_1|d)}{p(\mathcal{M}_2|d)} = \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)}B_{12} \qquad (5)$$

In other words, the prior odds of the models are updated by the data through the Bayes factor. If we do not have any special reason to prefer one model over the other before we see the data, then $\pi(\mathcal{M}_1) = \pi(\mathcal{M}_2) = 1/2$, and the posterior odds reduce to the Bayes factor. Alternatively, one can interpret the Bayes factor as the factor by which our relative belief in the two model is modified once we have seen the data.

B. A Bayesian measure of complexity

The usefulness of a Bayesian model selection approach based on the model likelihood is that it tells us whether the increased "complexity" of a model with more parameters is justified by the data. But the number of free parameters in a model is only the simplest possible notion of complexity, which we call *input complexity* and denote by $\mathcal{C}_0$. A much more powerful definition of model complexity was given by Spiegelhalter et al [6], who introduced the *Bayesian complexity*, which measures the number of model parameters that the data can constrain:

$$\mathcal{C}_b = -2\left(D_{KL}(p, \pi) - \widehat{D_{KL}}\right). \qquad (6)$$

On the right–hand side, $D_{KL}(p, \pi)$ is the Kullback-Leibler (KL) divergence between the posterior and the prior, representing the information gain obtained when upgrading the prior to the posterior via Bayes theorem:

$$D_{KL}(p, \pi) \equiv \int p(\theta|d) \log \frac{p(\theta|d)}{\pi(\theta)} d\theta \qquad (7)$$

The KL divergence measures the relative entropy between the two distributions (e.g. [4])[14]. In Eq. (6), $\widehat{D_{KL}}$ is a point–estimate for the KL divergence. If all parameters are well measured, then the posterior pdf will collapse into a small region around $\hat{\theta}$, and thus the KL divergence will approximately be given by $\widehat{D_{KL}} = \log p(\hat{\theta})/\pi(\hat{\theta})$. By taking the difference, we compare the effective information gain to the maximum information gain we can expect under the model, $\widehat{D_{KL}}$. The factor of 2 is chosen so that $\mathcal{C}_b \to \mathcal{C}_0$ for highly informative data, as shown below. As a point estimator for $\hat{\theta}$ we employ the posterior mean, denoted by an overbar. Other choices are possible, and we discuss this further below.

We can use Eq. (7) and Bayes theorem to rewrite (6) as

$$\mathcal{C}_b = -2\int p(\theta|d, \mathcal{M}) \log \mathcal{L}(\theta) + 2\log \mathcal{L}(\hat{\theta}), \qquad (8)$$

Defining an effective χ^2 through the likelihood as $\mathcal{L}(\theta) \propto \exp(-\chi^2/2)$ (any constant factors drop out of the difference of the logarithms in Eq. (6)) we can write the *effective number of parameters as*

$$\mathcal{C}_b = \overline{\chi^2(\theta)} - \chi^2(\hat{\theta}), \qquad (9)$$

where the mean is taken over the posterior pdf. This quantity can be computed fairly easily from a Markov chain Monte Carlo (MCMC) run, which is nowadays

widely used to perform the parameter inference step of the analysis.

We thus see that the effective number of parameters of the model is not an absolute quantity, but rather *a measure of the constraining power of the data as compared to the predictivity of the model, i.e.* the prior. Hence $\mathcal{C}_b$ depends both on the data at hand and on the prior available parameter space. In fact, it is clear that the very notion of "well measured parameter" is not absolute, but it depends on what our expectations are, *i.e.* on the prior. For example, consider a measurement of Ω_{tot}, the total energy density of the Universe, expressed in units of the critical density. The current posterior uncertainty around $\Omega_{\text{tot}} = 1$ is about 0.02. Whether this means that we have "measured" the Universe to be flat (*i.e.* $\Omega_{\text{tot}} = 1$) or not depends on the prediction of the model we consider. If we take a generic prior in the range $0 \leq \Omega_{\text{tot}} \leq 2$, then we conclude that current data have measured the Universe to be flat with moderate odds (a precise analysis gives odds of 18 : 1 in favor of the flat model, see [2]). On the contrary, in the framework of *e.g.* landscape theories, the prior range of the model is much narrower, say $|\Omega_{\text{tot}} - 1| \lesssim 10^{-5}$, and therefore current posterior knowledge is insufficient to deem the parameter measured.

Eq. (9) is conceptually related to the $\Delta\chi^2$ approach used for example in [10] to analyze how many dark energy parameters are measured by the data. In that work the decrease of the χ^2 value within the confidence regions encompassing 68% and 95% of the posterior was measured and compared against the theoretical decrease of a multivariate Gaussian distribution with a given number of degrees of freedom (see e.g. chapter 15.6 of [9]). This in turn allowed to deduce the effective number of degrees of freedom of the χ^2.

We now turn to an explicit demonstration of the power and usefulness of coupling the model likelihood with the Bayesian model complexity in model selection questions, by analyzing in some detail linear toy models.

III. LINEAR MODELS

Before applying the Bayesian complexity and model likelihood to current cosmological data, we compute them explicitly for a linear model and we illustrate their use in a toy example involving fitting a polynomial of unknown degree. We show below how the Bayesian complexity tells us how many parameters the data could in principle constrain given the prior expectations under the model.

A. Specification of the likelihood function

Let us consider the following *linear model*

$$y = F\theta + \epsilon \qquad (10)$$

where the dependent variable y is a d-dimensional vector of observations, θ is a vector of dimension n of unknown regression coefficients and F is a $d \times n$ matrix of known constants that specify the relation between the input variables θ and the dependent variables y [15]. Furthermore, ϵ is a d-dimensional vector of random variables with zero mean (the *noise*). If we assume that ϵ follows a multivariate Gaussian distribution with uncorrelated covariance matrix $C \equiv \text{diag}(\tau_1^2, \tau_2^2, \ldots, \tau_d^2)$, then the likelihood function takes the form

$$p(y|\theta) = \frac{1}{(2\pi)^{d/2} \prod_j \tau_j} \exp\left[-\frac{1}{2}(b - A\theta)^t(b - A\theta)\right],$$
$$(11)$$

where we have defined $A_{ij} = F_{ij}/\tau_i$ and $b_i = y_i/\tau_i$. This can be cast in the form

$$p(y|\theta) = \mathcal{L}_0 \exp\left[-\frac{1}{2}(\theta - \theta_0)^t L(\theta - \theta_0)\right], \qquad (12)$$

with the likelihood Fisher matrix L given by

$$L \equiv A^t A \qquad (13)$$

and a normalization constant

$$\mathcal{L}_0 \equiv \frac{1}{(2\pi)^{d/2} \prod_j \tau_j} \exp\left[-\frac{1}{2}(b - \theta_0 A)^t(b - A\theta_0)\right].$$
$$(14)$$

Here θ_0 denotes the parameter value that maximizes the likelihood, given by

$$\theta_0 = L^{-1} A^t b. \qquad (15)$$

As a shortcut, we will write

$$\chi^2(\theta) \equiv -2\ln p(y|\theta) = \chi^2(\theta_0) + (\theta - \theta_0)^t L(\theta - \theta_0), \quad (16)$$

where $\chi^2(\theta_0) = -2\ln\mathcal{L}_0$.

B. Model likelihood and complexity

In this section we gain some intuitive feeling about the functional dependence of the model likelihood and complexity on the prior and posterior for the simple case of linear models outlined above. The results are then applied in section III D to an explicit toy model, showing the model likelihood and complexity in action.

Assuming as a prior pdf a multinormal Gaussian distribution with zero mean and Fisher information matrix P (we remind the reader that the Fisher information matrix is the inverse of the covariance matrix), *i.e.*

$$\pi(\theta) = \frac{|P|^{1/2}}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\theta^t P\theta\right], \qquad (17)$$

where $|P|$ denotes the determinant of the matrix P, the model likelihood (3) and model complexity (9) of the linear model above are given by Eqs. (A3) and (A9), respectively (see Appendix A).

Let us now consider the explicit illustration of a model with n parameters, $\theta = (\theta_1, \ldots, \theta_n)$ and $\mathcal{C}_0 = n$. Without losing generality, we can always choose the units so that the prior Fisher matrix is the unity matrix, *i.e.*

$$P = \mathrm{Id}_n. \qquad (18)$$

This choice of units is natural since it is the prior that sets the scale of the problem. The likelihood Fisher matrix being a symmetric and positive matrix, it is characterized by $n(n+1)/2$ real numbers, which we choose to be its eigenvalues $1/\sigma_i^2$, $i = 1, \ldots, n$ and the elements of the orthogonal matrix U that diagonalizes it (corresponding to rotation angles). Here the σ_i represent the standard deviations of the likelihood covariance matrix along its eigendirections i, expressed in units of the prior width. With $D = \mathrm{diag}(\sigma_1^{-2}, \ldots, \sigma_n^{-2})$ we have that the likelihood Fisher matrix is given by $L = UDU^t$ and thus Eq. (A9) gives

$$
\begin{aligned}
\mathcal{C}_b &= \mathcal{C}_0 - \mathrm{tr}\left[\left(UDU^t + \mathrm{Id}_n\right)^{-1}\right] = \\
&= \mathcal{C}_0 - \mathrm{tr}\left[(D + \mathrm{Id}_n)^{-1}\right] = \sum_{i=1}^{n} \frac{1}{1 + \sigma_i^2}.
\end{aligned}
\qquad (19)
$$

The complexity only depends on how well we have measured the eigendirections of the likelihood function *with respect to the prior*. Every well–measured direction, (*i.e.*, one for which $\sigma_i \ll 1$) counts for one parameter in $\mathcal{C}_b$, while directions whose posterior is dominated by the prior, $\sigma_i \gg 1$, do not count towards the effective complexity. This automatically takes into account strong degeneracies among parameters. Notice also that once an eigendirection is well measured (*i.e.* in the limit $\sigma_i \ll 1$), then the prior width does not matter anymore.

In contrast, the model likelihood (A3) is given by (assuming for simplicity that the mean of the likelihood corresponds to the prior mean, *i.e.* $\theta_0 = 0$)

$$p(d|\mathcal{M}) = \mathcal{L}_0 \prod_{i=1}^{n} \frac{\sigma_i}{\sqrt{1 + \sigma_i^2}}. \qquad (20)$$

Finally, we remark that an important ingredient of the Bayesian complexity is the point estimator for the KL divergence. Here we adopt the posterior mean as an estimate, but other simple alternatives are certainly possible, for instance the posterior peak (or mode), or the posterior median. The choice of an optimal estimator is still matter of research (see *e.g.* section IV and the comments at the end of Ref. [6]). The important aspect is that the posterior pdf should be summarized by only one number, namely the value plugged into the KL estimator. This is obviously going to be a very bad description for highly complex pdf's, exhibiting for instance long, banana–shaped degeneracies. No single number can be expected to summarize accurately such a pdf. On the other hand, for fairly Gaussian pdf's all the different estimators (mean, median and peak) reduce to the same quantity. This clearly calls for using normal directions in

parameter space [12], which make the posterior as Gaussian as possible, a procedure that it would be wise to follow whenever possible for many other good reasons (*e.g.*, better and faster MCMC convergence).

C. Effective complexity as a data diagnostics

We now turn to the question of how we can use the model likelihood and complexity together as a tool for model selection. We shall see that the Bayesian complexity provides a way to assess the constraining power of the data with respect to the model at hand and to break the degeneracy among models with approximately equal model likelihood.

Let us consider two models A and B with different numbers of parameters, $n_B > n_A$ (but in general the two models need not to be nested). If the additional parameters of model B are required by the data, then the likelihood of model B will be larger and B will have larger posterior odds, thus it should be ranked higher in our preference than model A. However, even if the extra parameters of model B are not strictly necessary, they can lead to over–fitting of the data and compensate the Occam's penalty factor in (20) sufficiently to lead to a comparable marginal likelihood for both models. The effective complexity provides a way to break the degeneracy between the quality–of–fit term ($\mathcal{L}_0$) and the Occam's razor factor in the marginal likelihood, and enables us to establish whether the data is good enough to support the introduction of the additional parameters of model B.

To summarize, we are confronted with the following scenarios:

1. $p(d|B) \gg p(d|A)$: model B is clearly favored over model A and the increased number of parameters is justified by the data.

2. $p(d|B) \approx p(d|A)$ and $\mathcal{C}_b(B) > \mathcal{C}_b(A)$: the quality of the data is sufficient to measure the additional parameters of the more complicated model, but they do not improve its likelihood by much. We should prefer model A, with less parameters.

3. $p(d|B) \approx p(d|A)$ and $\mathcal{C}_b(B) \approx \mathcal{C}_b(A)$: both models have a comparable likelihood and the effective number of parameters is about the same. In this case the data is not good enough to measure the additional parameters of the more complicated model and we cannot draw any conclusions as to whether the additional complexity is warranted.

We illustrate these cases by computing the model likelihood and effective complexity of a toy model in the next section.

D. An illustrative example

As a specific example of the linear model described in section III A, we consider the classic problem of fitting data drawn from a polynomial of unknown degree. The models that we test against the data are a collection of polynomials of increasing order, with input complexity $\mathcal{C}_0 = n$, where n is the order of the polynomial. The question is then whether our model selection can correctly recover the order m of the polynomial from which the data are actually drawn.

The data covariance matrix is taken to be diagonal and with a common standard deviation for all points, s, while the prior over the polynomial coefficients is a multivariate Gaussian with covariance matrix given by the identity matrix. For definitiveness, we will take the underlying model from which the data are generated to have $m = 6$ parameters. First we draw $p = 10$ data points with noise $s = 1/200$. We plot in Figure 1 the resulting model likelihood and effective complexity as a function of the input model complexity, n. The likelihood of the models increases rapidly until $n = m$ and then decreases slowly, signaling that $n > 6$ parameters are not justified. (We plot the logarithm of the logarithm of the model likelihood as the models with $n < 6$ parameters are highly disfavored and would otherwise not fit onto the figure – an example of case (1) of the list in the previous section). The effective complexity on the other hand continues to grow until $n \sim p$, at which point the data is unable to constrain more complex models and C_b becomes constant. We conclude that the model with $n = 6$ is the one preferred by data and that additional parameters are not needed, although the data could have supported them. This is case (2) in the list of the previous section.

Next we decrease the number of data points to $p = 4$, in which case we obviously cannot recover more than four parameters. It comes as no surprise that the model likelihood stops increasing at $n = 4$, see Figure 2. But the effective complexity also flattens at $n = 4$, which means that the data cannot deal with more than four parameters, irrespective of the underlying model! In this case, corresponding to point (3) of the list in the previous section, we conclude that the available data do not support more than 4 effective parameters. On the other hand, we recognize that the flattening of the model likelihood at $n = 4$ does not necessarily mean that the underlying model has four parameters. We thus must hold our judgment until better data become available.

As an alternative to decreasing the number of data points, we can achieve a similar data degradation effect by keeping $p = 10$ data points but by increasing the noise to $s = 1$. We obtain a result similar to the previous case, which is plotted in Figure 3.

We conclude by emphasizing once more that in general the outcome of model selection based on assessing the model likelihood and effective complexity depends on the interplay of two factors. The first is the predictive power
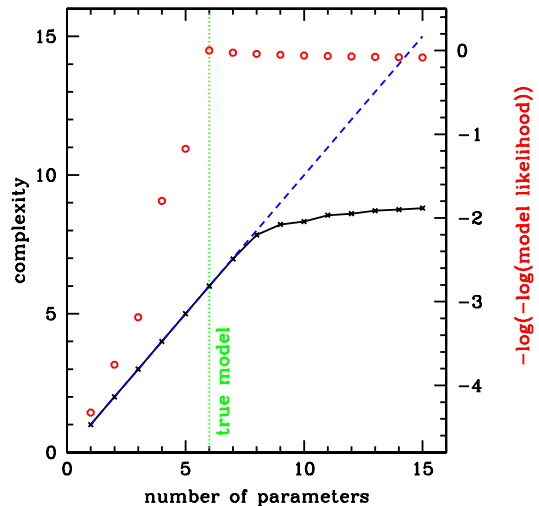

FIG. 1: Bayesian effective complexity C_b (solid black line, left–hand vertical scale) and model likelihood (red circles, right–hand scale) as a function of the number of parameters, for $d = 10$ data points with small noise. The dashed blue line is the number of parameters for reference. The errorbars on the model likelihood values are smaller than the symbols on this scale, while the Bayesian complexity is independent of the noise realization (i.e., error–free) for linear models. The Bayesian analysis correctly concludes that the best model is the one with $n = 6$ parameters.

of the model, as encoded in the prior. The second is the constraining power of the data.

IV. HOW MANY PARAMETERS DOES THE CMB NEED?

We now apply the above tools to the question of how many cosmological parameters are necessary to describe current cosmic microwave background (CMB) anisotropy measurements. We make use of the following CMB data: WMAP, ACBAR, CBI, VSA and Boomerang 2003. To provide an additional regularization (especially in view of including spatial curvature) we also use the HST limits on the Hubble parameter, $H_0 = 72 \pm 8$ km/s/Mpc. We should note that this strongly increases the power of the CMB data. We use both the first-year WMAP alone (WMAP1) as well as the data for the first three years (WMAP3).

For each set of cosmological parameters we create a converged MCMC chain using the publicly available code `cosmoMC` [11]. We then compute the Bayesian complexity from the chain through Eq. (9). The model likelihood is evaluated with the Savage-Dickey method (see [2] and references therein): For a model that is nested within a larger model by fixing one parameter, θ_f, to a value θ_0, the Bayes factor between the two models is given by the posterior of the larger model at $\theta_f = \theta_0$ (normal-
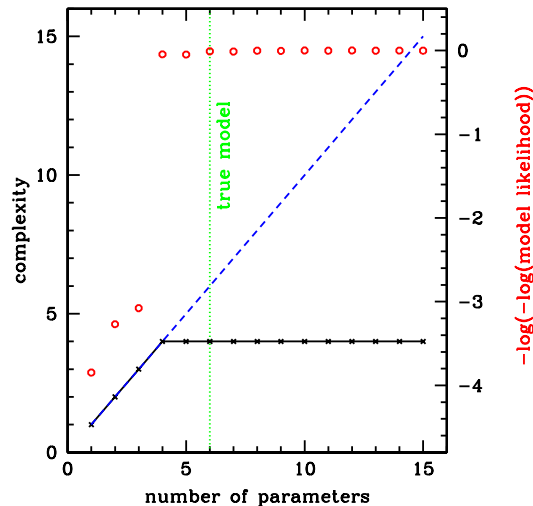
FIG. 2: As in Figure 1, but now using only $p = 4$ data points. The maximum effective complexity that the data can support is $\mathcal{C}_b = 4$, and the flattening of the model likelihood at the same value does not allow to conclude that models with more parameters are disfavored.
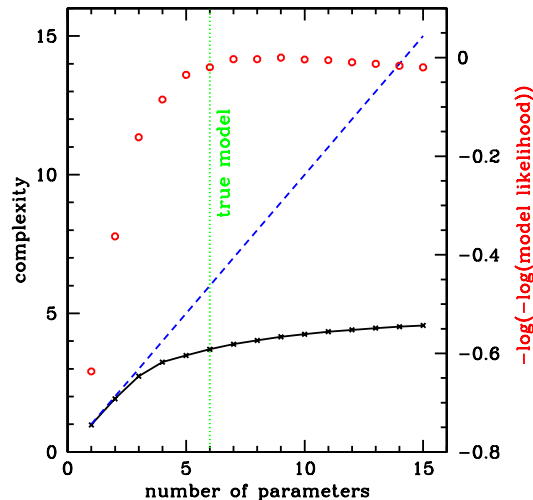

FIG. 3: As in Figure 1, but now with $p = 10$ data points and large noise. As in Figure 2, the maximum complexity supported by the data is smaller than the underlying true model complexity, $m = 6$.

ized and marginalized over all other parameters) divided by the prior at this point. In this way it is possible to derive all the model likelihoods in a hierarchy, starting from the most complex model (which is assigned an arbitrary model likelihood, in our case 1). Since errors tend to accumulate through the intermediate steps necessary to reach the simpler models, and as a cross–check, we additionally computed the model likelihoods with nested sampling [3] for the WMAP1 data. Within the errorbars,

we did not find any appreciable discrepancy between the two methods.

In this analysis we use four basic cosmological parameters, namely

$$b_4 = \{\Omega_b h^2, \Omega_m h^2, \theta_\star, A_s\}, \qquad (21)$$

where Ω_b (Ω_m) is the baryon (matter) density relative to the critical energy density, $h = H_0/100$ km/s/Mpc is the fudge factor, $\theta_\star$ is the ratio of the sound horizon to the last scattering angular diameter distance and $A_s = \ln P(k_0)$, with $P(k_0)$ the power spectrum of adiabatic density fluctuations at a scale $k_0 = 0.05$ Mpc^{-1}. We then add three more parameters in various combinations, to study whether they are necessary and supported by the observations. The additional parameters are: The reionization optical depth τ, the scalar spectral index n_s and the spatial curvature (parameterized by its contribution to the Hubble equation, Ω_K).

The scalar spectral index is either fixed to $n_s = 1$ (the case of a scale invariant power spectrum of initial fluctuations), or else chosen with a Gaussian prior, $n_s = 1 \pm 0.1$. We find that $\pi(n_s = 1) = 1/(0.1\sqrt{2\pi}) \approx 4$. We choose the prior of the reionization optical depth to reflect our current lack of understanding of how reionization proceeds. We choose it flat within $0 \leq \tau \leq 0.15$ and then add an exponential falloff:

$$\pi(\tau) \propto \exp\left(-\frac{\tau - 0.15}{0.05}\right) \text{ for } \tau > 0.15. \qquad (22)$$

The models without this parameter have $\tau = 0$, and $\pi(\tau = 0) = 5$. The curvature contribution is either set to zero, $\Omega_K = 0$, if we do not include the parameter, or else is used with a flat prior $-1 \leq \Omega_K \leq 1$. The value of the prior for a flat universe is $\pi(\Omega_K = 0) = 1/2$. We find that adding curvature as a free parameter when using the WMAP 3yr data leads to a non-Gaussian posterior for which the mean as a point-estimate for the KL divergence in Eq. (6) is not representative. We opt here for a slightly modified estimator, given by the average of the χ^2 evaluated at the mean and the mode of the posterior. For a Gaussian posterior this reduces to the mean point-estimator but it appears to be somewhat more stable.

We quote our results in Table I (WMAP3) and Table II (WMAP1), while Figure 4 gives a graphical representation. The model likelihoods are quoted relative to the model with the most parameters. We find that using WMAP1 we can measure all the parameters for the models with four and five parameters. For $\mathcal{C}_0 > 5$, the effective complexity increases more slowly than the number of input parameters, but we can still measure at least six parameters with CMB+HST. With WMAP3 we can measure all six parameters of the $b_4 + n_s + \tau$ model. We conclude that the new WMAP3 data augmented by the HST determination of H_0 can measure all seven parameters considered in this analysis.

Taking into account the model likelihood values, we find that the models b_4, $b_4 + \Omega_K$ and (to a lesser extent)

Model	Model likelihood	$\mathcal{C}_0$	Effective complexity	Comments
$b_4 + n_s + \Omega_K + \tau$	1	7	6.9 ± 0.3	Too many parameters
$b_4 + n_s + \Omega_K$	0.035 ± 0.005	6	6.3 ± 0.2	Ω_K disfavored
$b_4 + n_s + \tau$	48 ± 2	6	6.0 ± 0.04	$n_s + \tau$ favored
$b_4 + \Omega_K + \tau$	0.04 ± 0.01	6	6.4 ± 0.3	Ω_K disfavored
$b_4 + n_s$	2.2 ± 0.3	5	5.0 ± 0.04	n_s necessary
$b_4 + \Omega_K$	$(1.5 \pm 0.5) \times 10^{-5}$	5	4.8 ± 0.04	Ω_K disfavored
$b_4 + \tau$	3.5 ± 1	5	4.9 ± 0.04	τ necessary
b_4	$(1.5 \pm 0.5) \times 10^{-3}$	4	4.0 ± 0.04	Strongly disfavored

TABLE I: Relative model likelihood (normalized to the model with the most parameters) with WMAP 3yr data and effective Bayesian complexity for the models discussed in the text. $\mathcal{C}_0$ gives the number of parameters of the model. The error on the effective complexity was computed from random sub–chains and represents only the statistical error.

Model	Model likelihood	Effective complexity
$b_4 + n_s + \Omega_K + \tau$	1	6.2 ± 0.1
$b_4 + n_s + \Omega_K$	0.06 ± 0.01	6.0 ± 0.3
$b_4 + n_s + \tau$	42 ± 5	5.5 ± 0.02
$b_4 + \Omega_K + \tau$	0.68 ± 0.15	5.6 ± 0.2
$b_4 + n_s$	4.5 ± 0.9	4.9 ± 0.03
$b_4 + \Omega_K$	$(1.5 \pm 0.5) \times 10^{-4}$	5.0 ± 0.09
$b_4 + \tau$	17 ± 5	4.8 ± 0.03
b_4	$(2.3 \pm 0.6) \times 10^{-3}$	4.0 ± 0.05

TABLE II: Model likelihood and complexity as inferred from the first year WMAP data, for comparison with the values in table I. We see how the WMAP3 data increases the model likelihood of $b_4 + n_s + \tau$ relative to the simpler models $b_4 + \tau$ and $b_4 + n_s$.

$b_4 + n_s + \Omega_K$ as well as $b_4 + \tau + \Omega_K$ are strongly disfavoured. This shows that $\Omega_K = 0$ is preferred by current data, in agreement with the result of Ref. [2]. In general, adding in a non–zero spatial curvature leads to a well measurable decrease in the model probability that, together with the increase in effective model complexity, reinforces our belief that Ω_K can be safely neglected for the time being. Of course this result is partially a reflection of our choice of prior on Ω_K. However, it is important to remember that had we halved the range of this prior, the likelihoods for models with non–zero curvature would have only doubled. This would not change the results significantly. Alternatively, an inflation–motivated prior of the type $|\Omega_K| \ll 1$ would render the parameter unmeasured and irrelevant. In this case adding it would not change the effective complexity or the model likelihood at all.

We also find that the basic set b_4 must be augmented by either n_s or τ. The inclusion of both parameters at the same time was optional with the first year WMAP data only, but using WMAP3 we find that $b_4 + n_s + \tau$ has a significantly higher model likelihood than all other models investigated here. Also, where with WMAP1 we only gained half an effective parameter when going from $b_4 + \tau$ or $b_4 + n_s$ to $b_4 + n_s + \tau$, we now gain one full parameter. Thus we can now measure both parameters at the same time.

Overall, we conclude that $b_4 + \tau$ was a good and suffi-cient base model until a few months ago. Now $b_4 + n_s + \tau$ should be used. A wider prior on n_s would have only a minimal impact on the complexity of models including a tilt, since n_s seems to be rather well–measured when considered alone (ie, not in combination with τ). Inclusion of a non–vanishing curvature is discouraged by Bayesian model comparison. We find that a model with 6 parameters is sufficient to explain the current CMB data, even though all seven effective parameters can be constrained now. This analysis demonstrates that the Bayesian complexity estimator (9) works with real–world data and gives useful additional information for model comparison.

V. CONCLUSIONS

In this work we introduced the Bayesian complexity as a measure of the effective number of parameters in a model. We discussed extensively its properties and its usefulness in the context of linear models, where it can be computed analytically. We showed that it corresponds to the number of parameters for which the width of the posterior probability distribution is significantly narrower than the width of the prior probability distribution. These parameters can be considered to have been well measured by the data given our prior assumptions
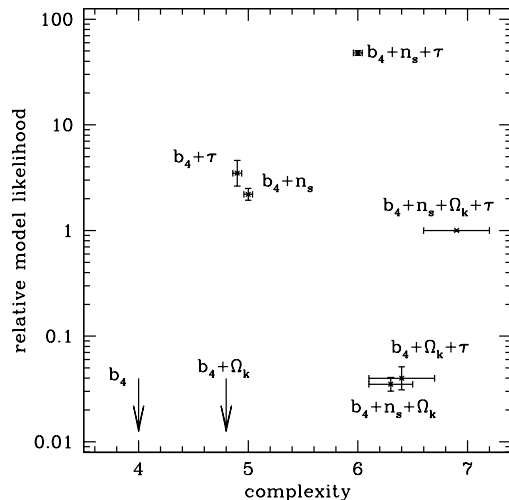
FIG. 4: We plot the model likelihood (normalized to the model with the most parameters) versus the Bayesian effective complexity for the models of Table I. A downward–pointing arrow indicates the Bayesian complexity of models that lie outside the boundary of the figure.

in the model.

We also showed that for linear models the Bayesian complexity probes the trace of the posterior covariance matrix, while the model likelihood is sensitive to the determinant. We argued that the Bayesian complexity allows to test for cases where the data is not informative enough for the model likelihood to be a reliable indicator of model performance, and provided an explicit example.

Finally, we applied the combination of model likelihood and Bayesian complexity to the question of how many (and which) parameters are measured by current CMB data, complemented by the HST limit on the Hubble parameter. We limited ourselves to the family of ΛCDM models with a power–law spectrum of primordial perturbations. We demonstrated that – in addition to the energy density in baryons and matter, the CMB peak location parameter $\theta_\star$ and the amplitude of the initial perturbations – we need to consider now both the reionisation optical depth and the scalar spectral index. Non–flat models are disfavoured. The effective complexity shows that the CMB data can measure all seven parameters in this scheme.

As the Bayesian complexity is very easy to compute from a MCMC chain, we hope that it will be used routinely in future data analyses in conjunction with the model likelihood for model building assessment. It will help to determine if the data is informative enough to measure the parameters under consideration. Further work is needed to study the performance of the Bayesian complexity in situations with a strongly non–Gaussian posterior.

APPENDIX A: MODEL LIKELIHOOD AND COMPLEXITY IN THE LINEAR CASE

Here we compute first the model likelihood (3) for the linear model (12), using the Gaussian prior (17). An analogous computation can be found in [6]. Returning to Bayes theorem (1), the posterior pdf is given by a multinormal Gaussian with Fisher information matrix F

$$F = L + P \qquad (A1)$$

and mean $\bar{\theta}$ given by

$$\bar{\theta} = F^{-1}L\theta_0. \qquad (A2)$$

The model likelihood (3) evaluates to

$$
\begin{aligned}
p(d|\mathcal{M}) &= \mathcal{L}_0 \frac{|F|^{-1/2}}{|P|^{-1/2}} \exp\left[-\frac{1}{2}\theta_0^t (L - LF^{-1}L)\theta_0 \right] \\
&= \mathcal{L}_0 \frac{|F|^{-1/2}}{|P|^{-1/2}} \exp\left[-\frac{1}{2}(\theta_0^t L\theta_0 - \bar{\theta}^t F\bar{\theta}) \right].
\end{aligned}
\qquad (A3)
$$

This can be easily interpreted by looking at its three components: the quality of fit of the model is encoded in $\mathcal{L}_0$, which represents the best–fit likelihood. Thus a model that fits the data better will be favored by this term. The term involving the determinants of P and F is a volume factor (the so called *Occam's factor*). As $|P| \leq |F|$, it penalizes models with a large volume of wasted parameter space, *i.e.* those for which the parameter space volume $|F|^{-1/2}$ that survives after arrival of the data is much smaller than the initially available parameter space under the model prior, $|P|^{-1/2}$. Finally, the exponential term suppresses the likelihood of models for which the parameters values that maximize the likelihood, θ_0, differ appreciably from the expectation value under the posterior, $\bar{\theta}$. Therefore when we consider a model with an increased number of parameters we see that its model likelihood will be larger only if the quality–of–fit increases enough to offset the penalizing effect of the Occam's factor.

Let us now turn to the computation of the Bayesian complexity, Eq. (9). Using the posterior mean (denoted by an overbar) as a point estimator for the effective chi–square, we obtain from (16)

$$\chi^2(\bar{\theta}) = \chi^2(\theta_0) + (\bar{\theta} - \theta_0)^t L(\bar{\theta} - \theta_0). \qquad (A4)$$

The expectation value of the χ^2 under the posterior is given by

$$\overline{\chi^2(\theta)} = \chi^2(\theta_0) + \text{tr}\left(L\langle(\theta - \theta_0)^t(\theta - \theta_0)\rangle\right). \quad (A5)$$

We concentrate on the second term and write $(\theta - \theta_0) = (\theta - \bar\theta) + (\bar\theta - \theta_0) = u + v$. The total term in the expectation value then becomes $\langle u^t u + u^t v + v^t u + v^t v\rangle$. The first term of this expression is just the posterior covariance matrix $\langle u^t u\rangle = F^{-1}$. The last term combines with $\chi^2(\theta_0)$ to $\chi^2(\bar\theta)$. The cross terms vanish since $\langle u\rangle = 0$.

All taken together, we obtain for the complexity

$$\mathcal{C}_b = \text{tr}\left(L\langle(\theta - \bar\theta)^t(\theta - \bar\theta)\rangle\right) \quad (A6)$$

$$= \text{tr}\left(LF^{-1}\right) \quad (A7)$$

Using the relation (A1) we can rewrite the complexity as

$$\mathcal{C}_b = \text{tr}\left\{(F - P)F^{-1}\right\} \quad (A8)$$

$$= \mathcal{C}_0 - \text{tr}\left\{PF^{-1}\right\}. \quad (A9)$$

Thus while the model likelihood depends on the determinant of the Fisher matrices, the complexity depends on their trace. Another important point worth highlighting is that for linear models the complexity does not depend on the degree of overlap between the prior and the posterior, nor on the noise realization (as long as the noise covariance matrix is known).

[1] A.R. Liddle, Mon. Not. Roy. Astron. Soc. **351**, L49-L53 (2004); T.D. Saini, J. Weller and S.L. Bridle, Mon. Not. Roy. Astron. Soc. **348**, 603 (2004); B.A. Bassett, P.S. Corasaniti and M. Kunz, Astrophys.J. **617**, L1 (2004); G. Lazarides, R.R. de Austri and R. Trotta, Phys. Rev. D **70**, 123527 (2004); A. Niarchou, A.H. Jaffe, L. Pogosian, Phys.Rev. D69 (2004) 063515; P. Mukherjee et al, astro-ph/0512484 (2005); M. Beltran et al, Phys. Rev. D **71**, 063532 (2005); M. Kunz et al, Phys. Rev. D **73**, 023511 (2006).

[2] R. Trotta, astro-ph/0504022 (2005).

[3] P. Mukherjee, D. Parkinson and A.R. Liddle, astro-ph/0508461 (2005).

[4] D.J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press (2003).

[5] C.E. Shannon, Bell Sys. Tech. J. **27**, 379 and 623 (1948).

[6] D.J. Spiegelhalter et al, J.R. Statist. Soc. B **64**, 583 (2002).

[7] D. Feldmann, Information Theory, Excess Entropy and Computational Mechanics, available from http://hornacek.coa.edu/dave/

[8] E.T. Jaynes and G.L. Bretthorst (ed.), *Probability Theory: The Logic of Science*, Cambridge University Press (2003).

[9] W.H. Press et al, *Numerical Recipes in C* (second edition), Cambridge University Press (1992).

[10] P.S. Corasaniti et al, Phys. Rev. D **70**, 083006 (2004).

[11] A. Lewis and S. Bridle, Phys.Rev. D **66**, 103511 (2002). The code is available from http://cosmologist.info

[12] A. Kosowsky, M. Milosavljevic and R. Jimenez, Phys. Rev. D **66**, 63007 (2002); M. Chu, M. Kaplinghat and L. Knox, Astrophys. J. **596**, 725 (2003).

[13] This is an effort to be consistent with common terminology in the statistical community and to avoid confusion with the colloquial meaning of the term "evidence".

[14] Notice that the KL divergence is not symmetric (*i.e.*, in general $D_{KL}(P, Q) \neq D_{KL}(Q, P)$), but it does have the useful properties that $D_{KL}(P, Q) \geq 0$ (Gibbs inequality), with equality only for $P = Q$, and that it is invariant under general transformations of the parameter space.

[15] In the special case of observations $y_i(x)$ fitted with a linear model $f(x) = \theta_j X^j(x)$, the matrix F is given by the basis functions X^j evaluated at the locations x_i of the observations, $F_{ij} = X^j(x_i)$.